Published in proceedings of 24th Text REtrieval Conference (TREC 2015), Gaithersburg, USA, 17-20 November 2015, which should be cited to refer this work.

Exploiting incoming and outgoing citations for improving Information Retrieval in the TREC 2015 Clinical Decision Support Track

J. Gobeill [ab], A. Gaudinat[a], P. Ruch[ab]

[a] *BiTeM group, University of Applied Sciences, Information Studies Department, Geneva*
[b] *SIBtex group, Swiss Institute of Bioinformatics, Geneva*

contact: {julien.gobeill;patrick.ruch}@hesge.ch

Abstract

We investigated two strategies for improving Information Retrieval thanks to incoming and outgoing citations. We first started from settings that worked last year and established a baseline. Then, we tried to rerank this run. The incoming citations' strategy was to compute the number of incoming citations in PubMed Central, and to boost the score of the articles that were the most cited. The outgoing citations' strategy was to promote the references of the retrieved documents. Unfortunately, no significant improvement from the baseline was observed.

Introduction

The text mining group at the Swiss Institute of Bioinformatics in Geneva has a long history of participation in TREC campaigns, including TREC Genomics [1], TREC Medical Records [2], TREC Chemical IR [3], and last year's TREC Clinical Decision Support track [4]. The group thus already works on medical Information Extraction from scientific publications, but also from social media. For example, 60,000 tweets are published daily about pharmaceutical drug specifications. As in 2014, the focus of the Clinical Decision Support Track was the retrieval of biomedical articles relevant for answering generic clinical questions about medical records.

Last year, we investigated a lot of strategies, such as sections indexing, Medical Subject Headings (MeSH) enrichment, or article-type boosting. This year we focused on re-ranking based on citations network. Basically, an article has two types of citations: incoming and outgoing citations. For example, if an article A contains a reference to an article B in its Reference section, we will say that article A has an outgoing citation to article B, and article B has an incoming citation from article A. Thus, both types of citations can be used in order to improve Information Retrieval. Such attempts are reported in the literature, from Salton in 1971 [5] to the complete thesis of Larsen [6].

For this campaign, we first started from the last year's best settings in order to establish a baseline. Then, we tried to improve this baseline with reranking strategies based on incoming, then outgoing citations.

Data and strategies

All retrievals were generated with Terrier [7], an IR platform (in Java) which implements state-of-the-art indexing and retrieval functionalities, including a TREC format output. In the following, the *Retrieval Status Value* (RSV) is the relevance score attributed to a document by Terrier. Post-processing strategies were applied thanks to local scripts in Perl.

1) Baseline

We used a linear combinations of two indexes with the Okapi BM25 and PL2 weighting schemes, with default settings, and an automatic relevance feedback query expansion: see [8] for more details about Terrier models.
In queries, all numbers were automatically discarded.

2) Exploiting incoming citations

For a document *d*, the number of incoming citations is the number of documents that contain a citation to this document *d*. Of course, this number grows with time. Moreover, it would be extremely difficult to locate all documents in the world, and to efficiently extract their references in order to precisely compute this number. Thus, we only worked with PubMed Central (PMC), in which references are well-structured. Even if we miss all non-PMC incoming citations, we assume that papers most cited in PMC are papers most cited in general. Then, reranking based on such popularity is feasible. PMC was accessed on July 2015.

For the TREC CDS documents, the average number of incoming citations from PMC was 3.2. Table 1 shows the decile analysis of this number.

Decile	Number of incoming citations
0%	0
10%	0
20%	0
30%	0
40%	0
50%	1
60%	2
70%	2
80%	4
90%	7
100%	2031

Table 1. Decile analysis of the number of incoming citations for the articles in the collection. Incoming citations were collected only from PubMed Central.

We thus used this number of citations (*nbcit*) in order to boost the score of the articles that are the most cited, thanks to Formula 1. *Alpha* is a setting parameter.

$$Score_d = RSV_d + \log(1 + nbcit_d) / alpha$$

Formula 1. Incoming citations boosting.

This boosting is a strict reranking according to the number of incoming citations: no new document is added in the results.

3) Exploiting outgoing citations

For exploiting outgoing citations, the idea was to promote the references of the retrieved documents. This strategy achieved leading results with patents (see TREC Chem campaigns [3], with up to +150% for MAP), but it was the first time we applied this to the medical literature. Formula 2 gives the final score of a document *d* after re-ranking. *E* is the set of retrieved documents (1000 by default), $is\_cited_{d,e}$ is 1 if document *d* is cited in document *e*, 0 otherwise. *Beta* is a setting parameter.

$$Score_d = RSV_d + \sum_E is\_cited_{d,e} \times beta \times RSV_e$$

Formula 2. Outgoing citations boosting.

This boosting is more than a strict reranking: new documents can appear, which were not retrieved by the search engine but cited by the retrieved articles.

Results and Discussion

In the following, we describe results in light of Top Precision, which is the Precision at interpolated Recall 0.

1) Boosting with incoming citations

Figure 1 shows results with different values of *alpha*, from 0.1 to 10. All results were computed after the competition.

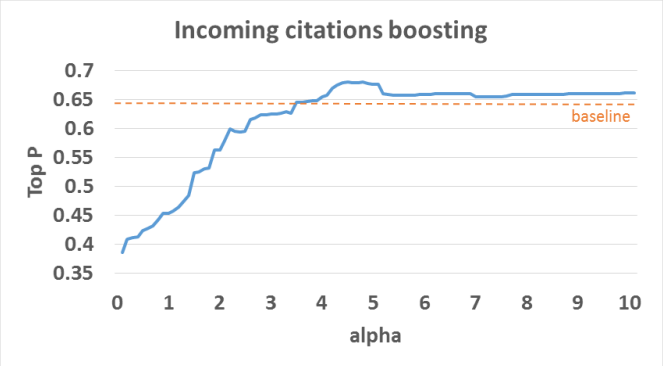

Figure 1. Top Precision with different values of alpha for the incoming citations boosting. Baseline was 0.64.

We observe a small optimum (TopP 0.66 with *alpha* between 4 and 5). Compared to the baseline (topP 0.645), the improvement is modest (+2.5%). Yet, the baseline R-Prec was never improved.

2) Boosting with outgoing citations

Figure 2 shows results with different values of *beta*, from 0.1 to 2. All results were computed after the competition.

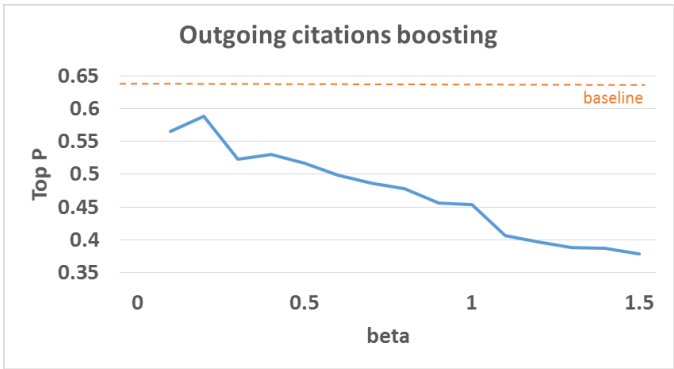

Figure 2. Top Precision with different values of beta for the outgoing citations boosting. Baseline was 0.64.

The best setting (beta 0.2) leads to a small decrease compared to the baseline (-10%).

Conclusion

We investigated two strategies for improving Information Retrieval thanks to incoming and outgoing citations. No improvement was observed.

References

[1] Gobeill, J., Tbahriti, I., Ehrler, F., & Ruch, P. (2007). Vocabulary-Driven Passage Retrieval for Question-Answering in Genomics. In Proceedings of the 2007 Text Retrieval Conference

[2] Gobeill, J., Gaudinat, A., Ruch, P., Pasche, E., Teodoro, D., & Vishnyakova, D. (2011). BiTeM Group Report for TREC Medical Records Track 2011. In Proceedings of the 2011 Text Retrieval Conference.

[3] Gobeill, J., Teodoro, D., Pasche, E., & Ruch, P. (2009). Report on the TREC 2009 Experiments: Chemical IR Track. In Proceedings of the 2009 Text Retrieval Conference.

[4] Gobeill, J., Gaudinat, A., Pasche, E., & Ruch, P. (2014). Full-texts representation with Medical Subjects Headings, and co-citations network reranking strategies for TREC 2014 Clinical Decision Support Track. In Proceedings of the 2014 Text Retrieval Conference

[5] Salton, G. (1971). Automatic indexing using bibliographic citations. Journal of Documentation, 27(2), 98-110.

[6] Larsen, B. (2004). References and citations in automatic indexing and retrieval systems-experiments with the boomerang effect (Doctoral dissertation, Det Informationsvidenskabelige AkademiDanish School of Library and Information Science, Institut østInstitut øst).

[7] Ounis I, Amati G, Plachouras V, et al. (2006) Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval. Terrier: A High Performance and Scalable Information Retrieval Platform.

[8] Amati G. (2009) Probabilistic Models for Information Retrieval based on Divergence from Randomness. Ph.D. thesis. Science University of Glasgow, Department of Computing.TREC-CHEM Track Guidelines.